Machine Learning – Supervised Learning

M-MLR-900
cyril.de-lajudie@epitech.eu
vincent.lemesle@epitech.eu
antoine.famibelle@epitech.eu

Part 1 & 2:

Beaucoup de mal à comprendre les consignes du a beaucoup de soucis notamment les lacunes en Math !

Part 3:

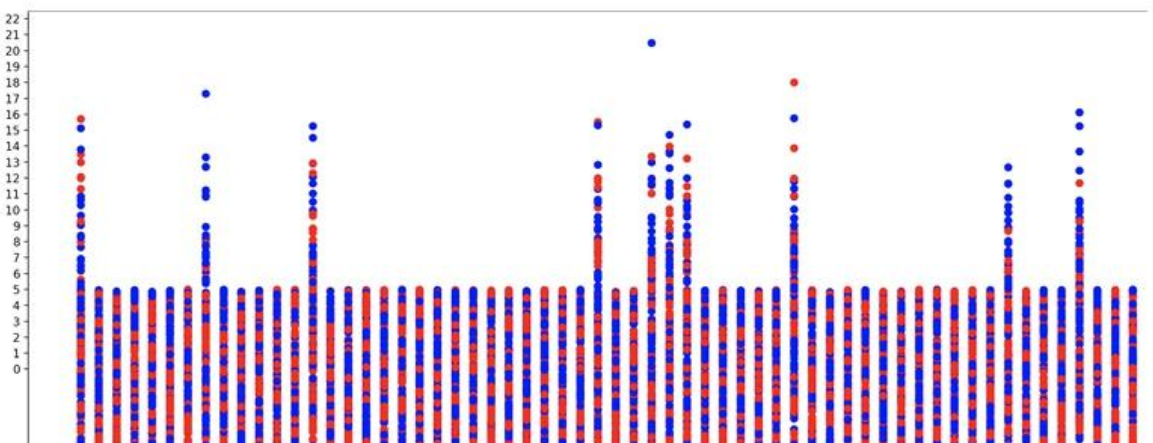
Observations

Afin de trouver une accuracy suffisante, nous avons decider d'utiliser différentes approches.

Tout d'abord, il nous semblait important de pouvoir visualiser l'ensemble de données.

Nous avons donc decider de creer un plot grâce à la librairie matplotlib. À partir de la nous avons decider de rajouter de la couleur à chacun des points afin de pouvoir gagner en lisibilité. Nous avons donc decider de mettre les données "gagnantes" en bleu et les autres en rouge.

Nous avons donc pu obtenir ce genre de visuel.



Avec les coordonnés en x afin de savoir de quel data il s'agissait.

Nous avons ensuite fait différents tests en utilisant et retirant certaines datèrent afin de voir comment variaient les différents algorithmes.

Nous les avons donc regroupé ou retiré en fonction de leurs variances visuelles.

Afin de ne pas régresser dans notre avancement nous avons ainsi établi un tableau. Dans ce dernier, nous notions la progression de l'accuracy pour nos quatre algorithmes définis au préalable. Nous avons établi un code couleur afin de nous représenter les différentes variations (vert = progression, jaune = aucun changement, rouge = régression).

Grâce à ce tableau, nous avons pu noter que certaines données étaient toujours utiles. Nous avons défini cela car elles permettent à un ou plusieurs algorithmes de progresser sans pour autant en faire régresser un autre.

	MODEL 1	MODEL 2	MODEL 3	MODEL 4
BASE SET	74.4	64.8	58.4	70.4
ADD 41	72.8	64.8	65.6	70.4
ADD 14	72.0	66.4	65.6	70.4
ADD 23	78.4	72.8	66.4	71.2
ADD 5	79.2	68.0	68.8	71.2
ADD 6	78.4	69.6	68.8	71.2
DLT 14	79.2	66.4	69.6	71.2
ADD 3	78.4	71.2	67.2	78.4
ADD 0	78.4	73.6	67.2	80.0
DLT 2	79.2	72.0	71.2	78.4
DLT 1	80.0	74.4	69.6	78.4
ADD 4	80.8	76.8	69.6	78.4
DLT 5	81.6	76.8	66.4	77.6
ADD 7	81.6	72.8	62.4	76.0
ADD 6	80.8	80.0	65.6	77.6
ADD 10	82.4	73.6	66.4	76.0
ADD 13	82.4	74.4	66.4	76.0
ADD 17	84.8	76.0	73.6	80.0
DLT 25	87.2	76.0	74.4	80.0

ADD 54	89.6	71.2	70.4	82.4
ADD 55	90.4	76.0	71.2	81.6
ADD 59	91.2	76.0	71.2	81.6

Conclusion

Pour la régression logistique, il existe différents tris de données qui permettent d'obtenir un résultat suffisant (> 85%). Nous avons reussi à atteindre un pic de 91.2% d'accuracy soit une accuracy supérieur à ce qui était demander (nombre d'erreurs réduites par 2).

Pour la KneighborsClassifier, il semblerait qu'il n'existe aucune combinaison ou tri de données permettant d'obtenir un résultat suffisant (maximum atteint 83%)

Pour la DecisionTreeClassifier, dû à son caractère aléatoire obéré lors de nos différents tests, nous avons prefere ne pas nous baser dessus afin d'avoir un résultat fiable et stable.

Pour le GaussianNB, il existe différents tris de données permettant d'avoir un résultat suffisant (accuracy > 85%). Cependant, l'accuracy obtenu grace à cet algorithme n'est jamais supérieur au meilleur résultat obtenu avec celui de la régression logistique. C'est pour cette raison que nous avons préférée opté pour un tri de données avec un algorithme de Régression logistique sur cette partie. Que ce soit d'un point de vue stabilité ou résultat cet algorithme convenait parfaitement a nos attentes.

Part 4:

Part 5:

Nous obtenons ces différentes accuracy à l'aide des différents classifier:

MODEL MEAN ACCURACY 93.00699300699301 %
KNeighborsClassifier ACCURACY 93.7062937062937 %
LinearDiscriminantAnalysis ACCURACY 97.2027972027972 %
GaussianNB ACCURACY 96.5034965034965 %
DecisionTreeClassifier ACCURACY 83.91608391608392 %
SVC ACCURACY 95.1048951048951 %